Research on Geolinguistic Linked Data: The Test Case of Cimbrian Varieties

Giorgio Maria Di Nunzio, Department of Information Engineering, University of Padua &
Stefan Rabanus, Chair of German Linguistics, Yerevan State Linguistic University

In this paper, we present a geolinguistic linked open data approach of a multidisciplinary and collaborative project, "Cimbrian as a test case for synchronic and diachronic language variation", which provides linguists with a test bed for formal hypotheses concerning human language. Aims of the project are to collect, digitize and tag linguistic data from the German dialect varieties of Cimbrian – spoken in three areas of northern Italy: Giazza (province of Verona), Luserna (province of Trento), and Roana (province of Vicenza) – and to make available on-line a valuable and innovative linguistic resource for the in-depth study of Cimbrian.

1 Introduction

Language resources that have been publicly made available can vary in the richness of the information they contain: on one hand, a corpus typically contains at least a sequence of words, sound or tags; on the other end, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody, and semantic content of every sentence, plus annotation of discourse relations or dialogue acts (cf. Bird/Klein/Loper 2009). When researchers need to perform particular linguistic analyses such as capturing fine-grained grammatical differences by comparing various dialectal translations of the same sentence, the only way to build a high accuracy language resource is by manual annotation (cf. Agosti et al. 2011, 63-64).

The heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections (cf. Chiarcos 2012). The rate of re-use for linguistic database technology together with related processing tools and environments is still too low. For example, the Edisyn search engine – the aim of which was to make different dialectal databases comparable – "in practice has proven to be unfeasible"[1] to date. In order to find common ground where linguistic material can be shared and re-used, the methodological and technological boundaries between different research projects have to be overcome.

The research direction we pursue in this work is to move the focus from the systems handling the linguistic data to the data themselves. We address these issues by adopting an approach based on the Linked Open Data (LOD) paradigm with the aim of enabling interoperability at a data-level by overcoming the characteristics of each collection which depend on different methodological and technological choices. For this purpose, we present a linguistic project which aims (i) to collect, digitize and tag linguistic data from the Cimbrian varieties and, (ii) to distribute data by means of an LOD. We also present a Web application which produces dynamic maps on user request that is built upon this open dataset.

[1] http://www.dialectsyntax.org/wiki/About_Edisyn. [All URLs in this paper were last accessed on January 17, 2013.]

2 Linguistic Project

In this contribution, we present the results of an ongoing multidisciplinary collaboration which is conducted in the context of the project named *Atlante Sintattico d'Italia*, Syntactic Atlas of Italy (ASIt)[2]. This project aims to implement a digital library system that provides access and enables management of curated dialect data, also by means of an advanced user interface specifically designed to update and annotate the linguistic data (cf. Agosti et al. 2012).

In this context, the Cimbrian project[3] focuses on the so-called Triveneto area in the north-eastern part of Italy, in which the Cimbrian dialects are in intense language contact with the Italian dialects belonging to the Lombard and Venetian dialect groups (cf. Pellegrini 1977). Cimbrian, spoken in the language island of Giazza (Veneto, province of Verona), Luserna (Trentino/South Tyrol, province of Trento) and – historically – Asiago/Roana (Veneto, province of Vicenza)[4], is of great interest to three important lines of research in linguistics:

- Romance dialectology: linguistic contact phenomena are visible especially at the lexical level,
- German dialectology: the language island varieties exhibit a high level of preservation of certain structural characteristics, and
- Historical linguistics: the diachronic development of a variety in isolation shows a particularly interesting mixture of preservation and innovation.

This historic language-contact situation (supplemented by the entry of spoken Regional Northern Italian in the repertoire of the speakers in the course of the 19th century) is crucial for our idea that language variation in Cimbrian depends both on its structural possibilities as a German dialect and on the multilinguism of its speakers. Hence, it is necessary to consider the Cimbrian and the Italian dialects of the area with respect to the same grammatical categories and features.

The interest for this linguistic context is witnessed by many studies on Cimbrian throughout the last decade (cf. the overviews in Bidese 2010). Furthermore, the present project, which puts its focus prominently on Cimbrian syntax, is coherent to similar projects at European level in that it creates a database of syntactic structures – which so far have been neglected in traditional dialectological work (cf. Rabanus/Alber/Tomaselli 2008). Finally, Cimbrian is an endangered language, with only few speakers of advanced age speaking Cimbrian fluently in Giazza[5]. This makes collection of linguistic data of this language all the more important.

[2] http://asit.maldura.unipd.it/.
[3] http://ims.dei.unipd.it/websites/cimbrian/.
[4] Additionally, some data from Mòcheno – another German-language island variety in Trentino which is collocated geographically and linguistically in between Cimbrian and Bavarian in South Tyrol (cf. Rabanus 2013) – have been considered. The entire area of Cimbrian and Mòcheno has been surveyed and documented in detail by Bruno Schweizer in the 1940's whose maps have been published as linguistic atlas (Schweizer 2012) only in the context of our project.
[5] The situation is much better in Luserna even though there are no children acquiring Cimbrian as mother language.

2.1 Documents

In contrast to many other German dialects Cimbrian has a tradition as written languages and a literature that goes back to the beginning of the 17th century. This makes it possible to reconstruct the language change for at least four empirically attested stages (1602, 1844, 1942, 2009/2010). The written documents that have been elaborated in order to form part of the database are "Christlike unt korze Dottrina" (1602, cf. Meid 1985), "Novena vun unzar liben Vraun" (1844, cf. Stefan 2000), "Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan" (1942, cf. Cappelletti/Schweizer 1942). These Cimbrian texts have been completely transcribed (faithfully to their graphic form) and segmented in sentences which have also been linked to their translations in Italian and Standard German. For contemporary Cimbrian fieldwork has been conducted in Giazza (2009 and 2010). In order to be able to compare the Cimbrian data with data from the Italian dialects and other projects on the syntax of German varieties, the questionnaire was designed as similar as possible to the ASIt questionnaires and has integrated questions elaborated by the SyHD project (*Syntax hessischer Dialekte*, Universities of Marburg/Frankfurt/Vienna)[6]. The interviews have been digitally recorded and transcribed both according to a Cimbrian orthography (developed for this purpose) and phonetically. The questionnaire so far aims to elicit syntactic and morphological data.

2.2 Tags

After segmentation of the sentences, tagging of the linguistic data is carried out. We start with tagging at the word-level, determining the parts of speech of single words. Tagging of syntactic phenomena at the sentence level and tagging of syntactic constituents will take place in a second phase of the project. The starting point for developing an adequate set of tags for Cimbrian is the tagset elaborated by the Edisyn project[7], especially for the (Dynamic) Syntactic Atlas of the Dutch dialects (DynaSAND)[8]. In collaboration with the ASIt team, we have developed a language-specific set of tags which is suitable for Cimbrian but, at the same time, allows the Cimbrian data to be linked to other databases of dialect syntax. This involves assigning the same names to same parts of speech as in the Edisyn and the ASIt databases, at most adding tags when they are needed for language-specific structures of Cimbrian, or leaving out tags which are not relevant for Cimbrian. Thus, for instance, the tag "verbal particle" has been added to identify verbal particles which can be found in German dialects (e.g. the verbal particle in the Standard German sentence, "Ich gehe weg", 'I go away'), but gender values such as "masculine" have been left out for the tag of the past participle, since past participles never inflect for gender in German varieties. We can therefore imagine the creation of a language-specific tagset as starting from a universal core, shared by all languages, and subsequently developing a language-specific periphery, which is compatible with other databases and appropriate to classify language-specific structures.

The sentences that are tagged can be searched by means of a search interface as shown in Figure 1 (see Section 3.2 for more details about this interface).

[6] http://www.syhd.info/.
[7] http://www.dialectsyntax.org/.
[8] http://www.meertens.knaw.nl/sand/.

3 Digital Geolinguistic Linked Open Data

The LOD paradigm refers to a set of best practices for publishing data on the Web[9] and it is based on a standardized data model, the Resource Description Framework (RDF).[10] RDF is designed to represent information in a minimally constraining way and it is based on the following building blocks: graph data model, URI-based vocabulary, data types, literals, and several serialization syntaxes.

3.1 Geolinguistic Ontology

The common ground defined by current European linguistic projects allows us to infer the fundamental classes and properties necessary to define an ontology for modeling and representing geolinguistic resources. Geolinguistic concepts can be organized into three major areas: geography, derivation, and tagging. The geographical area comprehends classes and properties related to physical places. The derivation area is about people speaking a certain language, their relationships, and the geographical area where they live. Furthermore, the derivation area allows for the study of the correlation between social factors, education and knowledge of the dialect, and the distinctiveness of a local dialect. Lastly, the tagging area regards language-specific classes and properties, such as documents, sentences, words, and their relationships (cf. Di Buccio/Di Nunzio/Silvello 2012, 2013a).

We present an example of how we built the ontology of a document. A document represents the composite unit of study of a dialect; it is composed by one or more sentences which are subsequently divided in words for further analysis. A document may be redacted in one language (e.g. Italian or English) and then translated into several dialects which allow for linguistic comparisons. The syntactical analyses of these parallel translations are possible thanks to the `Tag` class specialised into two main sub-classes: `Sentence Tag` and `PoS Tag`. `Sentence Tag` allows us to capture a sentence-level phenomena, whereas `PoS Tag` allows us to capture a phenomena occurring on a `Word` in a `Collocation`, i.e. a specific position, within a given `Sentence`. The `WordSentenceCollocation` class relates a tag to a word within a sentence along with the properties relating it to the `Word`, `Sentence` and `Collocation` classes. The `SentenceDocumentCollocation` class relates a sentence to a document specifying the collocation of the sentence within the document by means of the class `SentenceCollocation`.

This ontology is the starting point for modeling and describing geolinguistic resources because:

- it provides general-purpose concepts and relationships;
- it is extendable by adding more fine-grained classes;
- it permits an easy mapping from existing linguistic projects and publicly available databases.

[9] http://www.w3.org/DesignIssues/LinkedData.html.
[10] http://www.w3.org/RDF/.

This geolinguistic ontology allows us to expose the linguistic data as a Linked Open Dataset (see the details in Di Buccio/Di Nunzio/Silvello 2012a). Currently, the ASIt dataset is linked to DBpedia.[11]

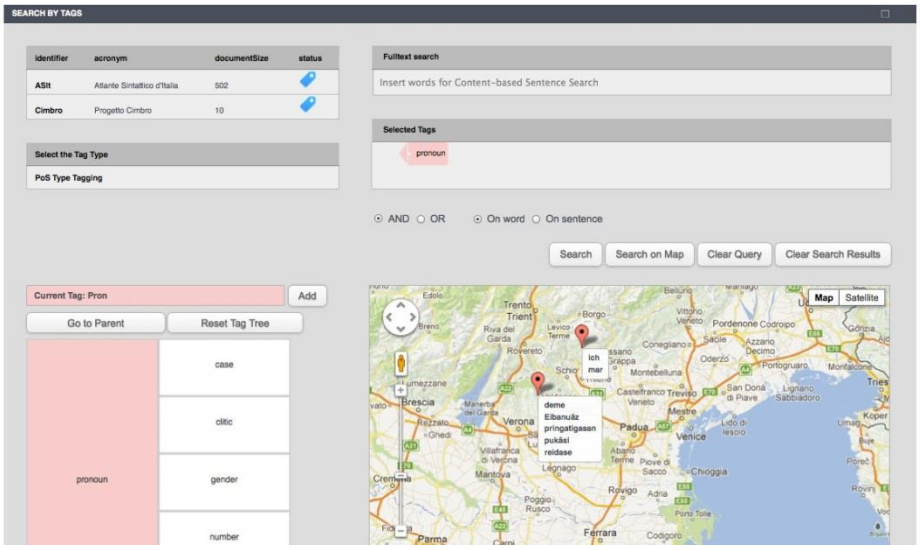

Figure 1: A screenshot of the ASIt GeoSearch interface.

3.2 Geolinguistic Web Application

The objective of the project is to provide linguists with a system for investigating variations among closely related languages. We developed a graphical user interface on top of the ASIt system that dynamically produces maps on the basis of the user request. The interface is available at the URL: http://svrims2.dei.unipd.it:8080/asit-enterprise/do/search.

A screenshot of a map produced by a tag-based search is reported in Figure 1. This type of search aims at satisfying the information need of a user searching for the geographic distribution of linguistic resources. The submitted query retrieves all the sentences that have the tags selected in the query ("pronoun" in this case). Then, for all these sentences, the system retrieves the locations and displays the words related to these tags (see details in Di Buccio/Di Nunzio/Silvello 2013).

3.3 Linguistic Analyses

The tagged corpus of Cimbrian data will be available to end users who might be linguists interested in carrying out syntactic analyses, or also informants, interested in correcting or augmenting the data. Concerning the former, it is important that the data are presented in a way which makes it usable by linguists working in different theoretical frameworks. Although it is inevitable (and, to some extent, also desirable) that the tagging of the data is influenced by theoretical considerations (in our case, the framework of genera-

[11] http://www.dbpedia.org/.

tive linguistics), it is important that the database should be of use not only to a small group of specialists.

With respect to the types of structures which can be analyzed in the tagged Cimbrian database, it will be possible to analyze syntactic structures and phenomena in great detail. It should also be possible to deduct morphological paradigms without too much effort, while it still remains a desideratum of further research projects to integrate a component which will make it possible to carry out phonological analyses on the database.

It is important that the structures in the database can be compared with structures present in other databases, since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, which is in contact with Romance varieties (hence can be compared to the ASIt data) but has a Germanic base (hence can be compared, e.g., to the DynaSAND data). To make just one example of what an analysis in these terms could look like, consider the case of pronouns and clitics in Cimbrian. In Cimbrian documents, sentences as the following can be found (Bidese 2008, p. 134):

miar	importar-z-mar	nicht	zo	sterben
me	matter-it-me	not	to	die

'I don't mind dying'

Whereas the use of the infinitive particle *zo* and the expletive pronoun *–z* are typical of German varieties, the doubling of the object pronoun *miar, mar* could be evidence for the development of a Romance-like system of clitics in Cimbrian, differently from Standard German where clitics are not attested. The tagged database will make it possible to retrieve all sentences of the corpus containing potential clitics and will therefore create an empirical basis on which to test hypotheses as those of the development of a system of clitics in Cimbrian.

4 Conclusions

In this paper, we presented the results of an ongoing linguistic project which aims to collect, digitize and tag linguistic data from the German dialect varieties of Cimbrian. The project gave the opportunity to merge different fields of research and begin a multidisciplinary collaboration between linguists and computer scientists. Since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, the main aim was to design and implement a digital library system that enables the management of linguistic resources of curated dialect data and provides access to grammatical data by means of a LOD approach. We imagine the use of the Geolinguistic Linked Open Dataset by third-party linguistic projects in order to enrich the data and build-up new services over them. To this purpose, we developed a graphical user interface on top of these linked data that dynamically produces maps on the basis of the user requests.

5 Acknowledgements

This work has been supported by the Project FIRB "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB Futuro

in ricerca 2008, cod. RBFR08KRA 003). We would like to thank Maristella Agosti, Emanuele Di Buccio, and Gianmaria Silvello of the Department of Information Engineering of the University of Padua, Paola Benincà and Diego Pescarini from the Department of Linguistic and Literary Studies of the University of Padua, Alessandra Tomaselli and Birgit Alber from the Department of Foreign Languages and Literatures of the University of Verona.

References

Agosti, M. et al. (2011): "A Digital Library of Grammatical Resources for European Dialects", in: Agosti, M. et al. (eds.): *Digital Libraries and Archives. 7th Italian Research Conference, IRCDL 2011. Pisa, Italy, January 20-21, 2011. Revised Selected Papers,* Berlin, Heidelberg, 61-74.

Agosti, M. et al. (2012): "A curated database for linguistic research: The test case of cimbrian varieties", in: Choukri, K. et al. (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 23-25. European Language Resources Association (ELRA)*, 2230-2236.

Bidese, E. (2008): *Die diachronische Syntax des Zimbrischen*, Tübingen.

Bidese, E. (ed.) (2010): *Il cimbro negli studi di linguistica*, Padua.

Bird, S./Klein, E./Loper, E. (2009): *Natural Language Processing with Python*, Sebastopol.

Di Buccio, E./Di Nunzio, G./Silvello, G. (2012): "A system for exposing linguistic linked open data", in: *Research and Advanced Technology for Digital Libraries International Conference on Theory and Practice of Digital Libraries (TPDL 2012), Papho, Cyprus, September 23-27,* Berlin, Heidelberg, 172–178.

Di Buccio, E./Di Nunzio/G., Silvello, G. (2013a): "A curated and evolving linguistic linked dataset", in: *Semantic Web*, 4, 3, 265-270.

Di Buccio, E./Di Nunzio, G./Silvello, G. (2013b): "A geolinguistic web application based on linked open data", in: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, 1101-1102.

Cappelletti, G./Schweizer, B. (1942): *Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan*, Bolzano.

Chiarcos, C. (2012): "Interoperability of corpora and annotations", in: Chiarcos, C./Nordhoff, S./Hellmann, S. (eds.): *Linked Data in Linguistics*, Berlin, Heidelberg, 161–179.

Meid, W. (1985): *Der erste zimbrische Katechismus*. Christlike unt korze Dottrina. *Die zimbrische Version aus dem Jahre 1602 der* Dottrina Christiana Breve *des Kardinal Bellarmin in kritischer Ausgabe. Einleitung, italienischer und zimbrischer Text, Übersetzung, Kommentar, Reproduktionen*, Innsbruck.

Pellegrini, G. (1977): *Carta dei dialetti d'Italia*, Pisa.

Rabanus, S./Alber, B./Tomaselli, A. (2008): „Erster Veroneser Workshop ‚Neue Tendenzen in der deutschen Dialektologie: Morphologie und Syntax'", in: *Vorschläge für die Ausrichtung zukünftiger Dialektsyntaxprojekte. Zeitschrift für Dialektologie und Linguistik,* 75, 72–82.

Rabanus, S. (2013): "La cartografia linguistica del mòcheno", in: Bidese, E./Cognola, F. (eds.): *Introduzione alla linguistica del mòcheno*, Turin, 129-146.

Schweizer, B. (2012): *Zimbrischer und Fersentalerischer Sprachatlas/Atlante linguistico cimbro e mòcheno*. Edited and commented by S. Rabanus, Luserna, Palù del Fersina.

Stefan, B. (2000): Novena vun unzar liben Vraun. *Die Zimbrische Mariennovene des D. Giuseppe Strazzabosco mit Übersetzung und Kommentar*, Innsbruck.